Research Paper | Education | E-ISSN No : 2454-9916 | Volume : 10 | Issue : 1 | Jan 2024

AN ANALYTICAL REVIEW OF BIG DATA AND HADOOP INTEGRATION

Maitri Rajesh Gandhi
Bhuj, Kachchh

ABSTRACT

The integration of Big Data with Hadoop represents a major advancement in the field of data management and analytics. This analytical review's goal is to examine the intricacies of this integration by examining its applicability, challenges, and potential applications. This article aims to shed light on key ideas, strategies, and new developments in order to present a comprehensive picture of the field. This is achieved by synthesizing recent research and viewpoints from seasoned business experts. It illuminates the revolutionary impact that the combination of Big Data and Hadoop has had on a range of sectors via critical analysis and empirical data, paving the path for better decision-making, innovation, and competitive advantage.

KEYWORDS: Bigdata, Hadoop, Mapreduce, Hive, Pig, HBAS

INTRODUCTION

Modern technology is now required for data processing, analysis, and storage since the volume of data produced in the last several years has reached previously unheard-of levels. Big Data presents opportunities as well as challenges for businesses operating in a variety of industries. It is characterized by its volume, velocity, and diversity. A prominent framework for distributed storage and processing of large information, the Hadoop framework has developed in response to the deluge of newly created data. There is a lot of promise when Big Data and Hadoop are combined since organizations will be able to find patterns, make informed decisions, and get actionable insights. In order to offer an analytical evaluation of the integration, this article will look at how Big Data and Hadoop have affected businesses, schools, and society overall. Our goal is to identify the most important parts of this integration by conducting a comprehensive literature review. We will also propose possible areas for future research and evaluate how well this integration is working. The potential for Big Data and Hadoop to operate together is better understood thanks to this study, which integrates theoretical frameworks with real research and practical insights. Consequently, it stimulates creativity and informs strategic endeavors.

Characteristics of big data:

The properties of big data are mostly comprised of the 3 V's.

1) Volume:

Additionally, the volume includes the total quantity of data that was produced by the organization. A data set's size is the determining factor in determining whether or not it is regarded to be big data.

2) Variety:

The nature of the data, the source of the data, and the type of data are all defined by the variety of data. The individuals that analyze the data are able to better understand the structured, unstructured, and semi-structured data as a result of this.

3) Velocity:

Data velocity refers to the pace of data creation and processing in reaction to demand. To put it another way, it refers to the flow of information.

Big data Architecture:

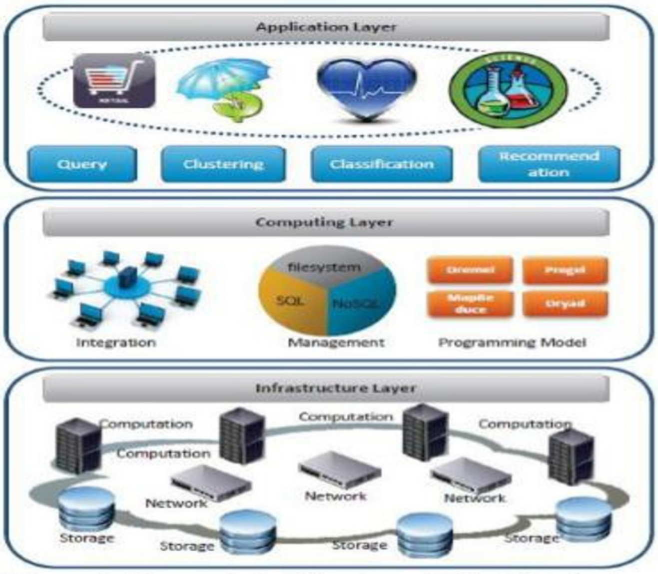

Figure 1: Layered Architecture of Big Data System

Problem associated with big data processing:

1) Information growth:"

In the big data realm, data size matters a great deal. What immediately comes to mind, even if we are familiar with the term "big data," is the sheer volume of data. The handling of a large amount of data that is expanding exponentially is the

challenging issue at hand. The CPU's performance is staying the same, but the volume of data is expanding faster.

2) Speed:

Size matters when it comes to speed. The size of the dataset and the amount of information it includes will determine how long it takes to answer.

3) Privacy And Security:

When dealing with massive amounts of data, the most pressing concern is data privacy. The inappropriate use of personal data is a major source of worry in the United States.

Hadoop-solution to the big data:

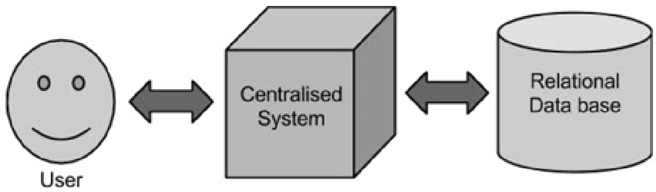

Figure 2. Actual working of hadoop

It has been shown that the above-described method works well with the application that handles a small volume of data. However, it is inappropriate for an application where a significant amount of data needs to be handled.

Google created Hadoop as a technology approach to deal with this problem. Using the mapReduce technique, which divides a task into smaller components and assigns each to a separate computer, Hadoop does this procedure. It then gathers the results so that they can be added to the dataset. The MapReduce algorithm is employed when applications are run on Hadoop. With this approach, the data is processed concurrently with other programs. In summary, Hadoop is used to create systems that can process enormous amounts of data and do in-depth statistical analysis.

Hadoop Architecture:

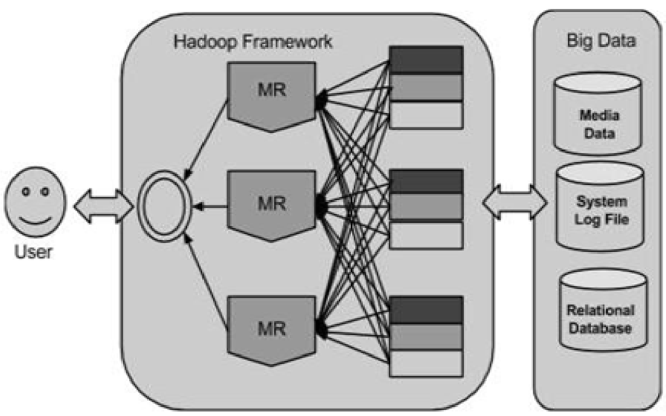

Figure 3. Hadoop Architecture Hadoop Architecture:

"Distributed computing makes use of a programming framework known as Hadoop to facilitate the processing of massive data sets. A software framework called Hadoop was developed by Google's MapReduce. It enables an application to be divided into many pieces. Hadoop utilizes two primary:

(a) Storage layer (Hadoop Distributed File System).
(b) Processing layer (MapReduce)"

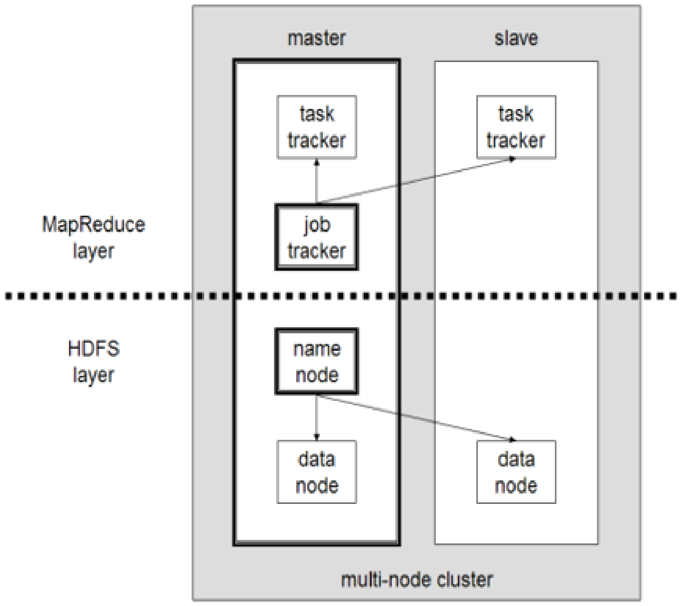

Figure 4. Hadoop Architecture

a) HDFS (Hadoop Distributed File System):

One such distributed file system is Hadoop's Distributed File System (HDFS), which takes its cues from Google's File System (GFS). Massive amounts of data can be stored on HDFS, which can also grow gradually and withstand the loss of data in the event that important parts of the storage infrastructure fail. Large volumes of data may be stored on HDFS, which also provides more convenient access

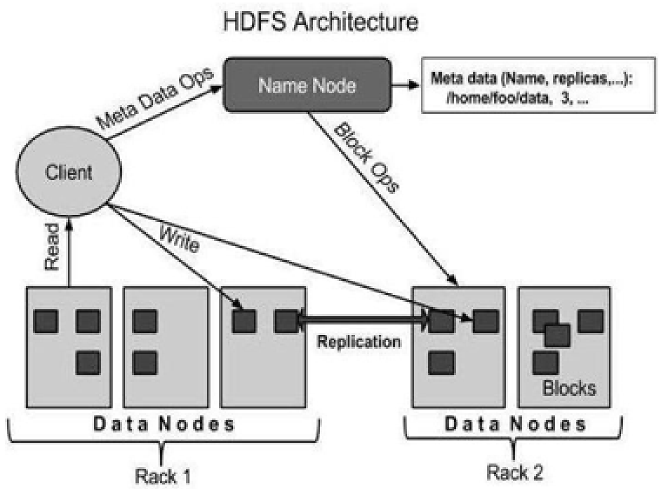

Figure 5. HDFC Architecture

b) MapReduce Architecture:

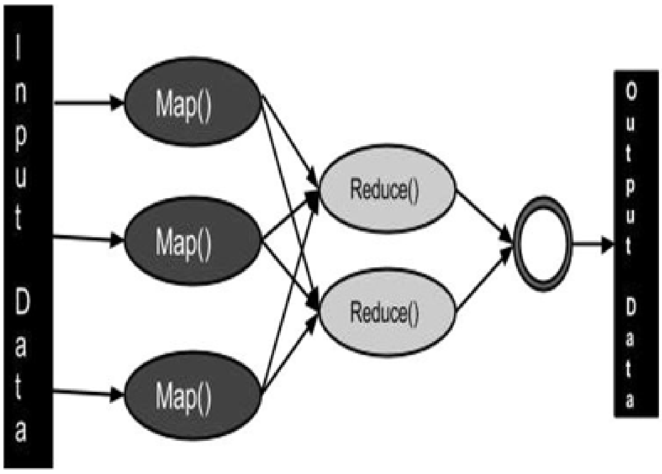

Figure 6. MapReduce Architecture

The MapReduce architecture serves as the foundation for Hadoop and is the processing column. The framework receives a massive amount of data, which is then divided into sections and run concurrently. The two essential tasks that make up the MapReduce algorithm are called Map and Reduce, respectively. The map function is used to transform a collection of data into another set of data. The individual elements in this new data collection are divided into tuples, or pairs of keys and values. The second work is called a reduction task, which takes an input consisting of a map's output and combines the data tuples related to that map into a more comprehensible set of records.

The MapReduce algorithm consists of three phases: mapping, shuffling, and reducing.

1. Map stage: It is the responsibility of the map or mapper to process the given data. It is common practice to save the input data in a file or directory format that may be stored in the Hadoop file system (HDFS). The input file is received line by line when it is given to the mapper function. After the data is processed, the mapper produces several extremely small data pieces.
2. Reduce stage: This level is the outcome of the combination and is a hybrid of the Reduce and Shuffle stages. The onus for handling the mapper-provided data falls on the Reducer. Processing results in a fresh set of outputs, which are then saved to the HDFS.

Other components of Hadoop:

Hive

The way that data is distributed and arranged in Hadoop for searching and organization is managed by the data warehousing program called Hive. Hadoop makes use of Hive, another well-liked programming environment, to enable the writing of data queries. Hadoop applications are written in declarative languages like Hive, however these languages do not allow for real-time queries. Hadoop may be transformed into a data warehouse with the help of Facebook's Hive technology. It also has a SQL dialect that might be helpful when processing queries. One example of a declarative language is Hive, which is a SQL dialect. Conversely, we tell Hive what we want, and it finds out how to construct a data flow to acquire it. "You are the one who gets to decide how data flows. Hive, in contrast to Pig, does require a schema, however you are not confined to just one. [2.0] Similar to PigLatin and SQL, Hive is a relationally complete language; however, it lacks Turing completeness. Even better, it can be extended to a Turing completeness with the help of UDFs like Piglatin. With the help of Hive, Hadoop can be transformed into a data warehouse." A SQL dialect is also included for searching data. Hive relies on tables to function. You can create two types of tables: managed tables, where Hive handles the data management, and external tables, where a third party outside of Hive handles the data management.

Pig

Pig is a procedural language that was created expressly to be used in the Hadoop environment for building large-scale parallel processing systems. Pig is a MapReduce substitute that uses its application to automatically construct MapReduce routines. Pig Latin is one of the components of the scripting language Pig. Pig uses Pig Latin as a translation tool for languages written in MapReduce. Pig is made up of two components: a language and an execution environment. PigLatin is the data flow language used by Pig. You program using this language by making links between various pieces. Piglet is another name for Pig's language. Pig is capable of handling operations on large data structures, even ones with several nesting levels. Pig is a database management system that does not require a schema for the data, in contrast to SQL. It is therefore more suited for handling unstructured data. However, if you would want to provide a schema, Pig may still utilize its value. Not only is PigLatin more complex than relational algebra, but it is also relationally complete, much like SQL. To achieve Turing completeness, one must use looping constructs, an unlimited memory model, and conditional structures. Adding User--Function That Is Defined can transform PigLatin into a Turing-complete language, even if it isn't one already.

HBase

Here we have HBase, a NoSQL database that is both scalable and distributed. Tables containing structured data may have an infinite number of rows and columns, as this was one of its primary design goals. Since it is not a relational database, HBase was never intended to handle real-time applications or any other type of transaction. Distributed column-based database type layer Apache HBase is built on top of Hadoop. The daily capacity was supposed to be billions of messages. The scalability of HBase is second to none, and it also provides fast random writes and random and streaming reads. It ensures atomicity at the row level as well, but it doesn't handle cross-row transactions natively. Big rows enable the creation of billions of indexed objects in a single database, and users have a lot of freedom when storing data in a column-oriented data structure. HBase is well-suited for tasks that need massive data storage, rapid scalability, and extensive indexing. Write-intensive workloads are a perfect fit for HBase.

Advantages and disadvantages of Hadoop: Advantages:

1) Range of data sources:

The data's structure will depend on the kind of information it

receives from various sources. Sources can also be found in online conversations, such as emails and social media posts. It would be quite time-consuming to transform all of the collected data into a single format. Hadoop is efficient as it can glean useful insights from any data type. In addition, it serves several purposes, including data warehousing and fraud detection.

2) Cost effective:
Companies have to set aside a significant portion of their profits for data storage because of the sheer amount of data. On occasion, they had to remove a large quantity of raw data in order to accommodate fresh data. There was a risk of losing important data because of the conditions. Hadoop was important in solving this issue thoroughly. It is a cost-effective and efficient way to store data.

3) Speed:
Any company worth its salt employs some kind of platform to streamline its operations. Hadoop helps the company fulfill its data storage needs in this way, among others. It stores data via a distributed file system, which involves storing it in several locations.

4) Multiple copies:
Hadoop automatically replicates and stores several copies of the data created by the programming language. The data will remain intact even if there is a very improbable failure, which is why all of this is done. Because Hadoop knows how important the organization's data is, it won't remove it unless the company says so.

Disadvantages:
1) Lack of preventive measures:
It is essential to provide the necessary safety precautions while working with sensitive data that has been collected by a company. By default, the security safeguards are disabled while utilizing Hadoop. It is incumbent upon the person in charge of the data to be cognizant of this and to take the appropriate steps to guarantee the security of the data.

2) Small Data concerns:
Certain big data systems that are already on the market are not appropriate for uses involving small amounts of data. Hadoop is an example of a platform whose characteristics are restricted to use by large enterprises producing vast volumes of data. It cannot function well in situations where there is little data.

3) Risky functioning
One of the most popular programming languages at the moment is Java. Furthermore, because hackers can easily take advantage of Java-based frameworks, it has been associated with several groups. Hadoop is only one example of a system of that type; it is built completely of Java. As a result, the platform is prone to degradation, which might have unforeseen repercussions.

Applications:
1) Amazon:
Clusters can include one to one hundred nodes in order to build Amazon's product search indexes and analyze millions of sessions per day for analytics using both the Java and streaming APIs.

2) Yahoo!:
Web search and ad systems research benefit from the usage of Hadoop. About 20,000 computers have more than 100,000 central processing units (CPUs). Every node in the biggest cluster, which has 2000 of them, has a 4TB disk.

3) Facebook:
It takes around 1.3 PB of raw storage and 320 machines to run a 320-machine cluster with 2,560 cores for maintaining copies of internal log and dimension data sources that are used for machine learning, reporting, and analytics.

2) CONCLUSION
Big Data and Hadoop represent a dramatic change in the manner that data management and analytics are usually done. Hadoop has democratized access to Big Data technologies, enabling businesses of all sizes to take advantage of the potential of data-driven insights. Hadoop's scalable architecture, fault tolerance, and affordability make this feasible. However, there are several challenges with this integration, such as those related to data security, privacy, and quality. It is imperative that stakeholders take proactive steps to address these challenges and maximize the tremendous potential of Big Data and Hadoop integration as they navigate this complex landscape.

Further research is required in the future to examine emerging patterns in the Big Data and Hadoop environment. Real-time analytics, edge computing, and the incorporation of machine learning apps are a few instances of these developments. Promoting interdisciplinary collaboration and knowledge sharing can help open up new avenues for innovation and propel sustainable growth in the digital age. Big Data and Hadoop integration success will mostly depend on three factors: organizational agility, strategy alignment, and a commitment to data-driven excellence.

REFERENCES

1. Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar," A Review Paper on Big Data and Hadoop", International Journal of Scientific and Research Publications, 10, October 2014.
2. Yashika Verma, Sumit Hooda," A Review Paper on Big Data and Hadoop", International Journal of Scientific and Research Publications, Feb,2015.
3. Iqbaldeep Kaur, Navneet Kaur, Amandeep Ummat, Jaspreet Kaur, Navjot Kaur," Research Paper on Big Data and Hadoop", International Journal Of Computer Science And Technology publicatios, Oct-Dec 2016.
4. Tom White," Hadoop: The Definitive Guide", O'Reilly Media Publication ,3rd Edition
5. Garry Turkington," Hadoop: Beginner's Guide", Packt Publishing,2013
6. http://www.tutorialpoints.com/hadoop _overview.pdf
7. http://blogs.mindsmapped.com/bigdatahadoop
8. https://www.knowledgehut.com/blog/bigdata-hadoop/top-pros-and-cons-of-hadoop
9. https://readwrite.com/2013/05/23/hadoop/applications
10. https://datajobs.com/what-is-hadoop-and-nosql
11. https://www.quora.com/What-is-the-difference-between-HBase-

and-Hadoop
12. https://www.sciencedirect.com/science/article/pii/S221457961730014X#se0160